Guest Editorial: Image and Video Inpainting and Denoising

Sergio Escalera, Hugo Jair Escalante, Xavier Baró, Isabelle Guyon, Meysam Madadi, Jun Wan, Stéphane Ayache, Yağmur Güçlütürk, Umut Güçlü

Guest Editorial: Image and Video Inpainting and Denoising

Article *in* IEEE Transactions on Pattern Analysis and Machine Intelligence · May 2020

DOI: 10.1109/TPAMI.2020.2971291

9 authors, including:

Sergio Escalera
University of Barcelona
350 PUBLICATIONS **4,781** CITATIONS
SEE PROFILE

Hugo Jair Escalante
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)
231 PUBLICATIONS **3,051** CITATIONS
SEE PROFILE

Xavier Baró
Universitat Oberta de Catalunya
110 PUBLICATIONS **1,805** CITATIONS
SEE PROFILE

Isabelle Guyon
Université Paris-Saclay and INRIA
209 PUBLICATIONS **37,240** CITATIONS
SEE PROFILE

Guest Editorial: Image and Video Inpainting and Denoising

Sergio Escalera, Hugo Jair Escalante, Xavier Baró, Isabelle Guyon, Meysam Madadi, Jun Wan, *Senior Member, IEEE*, Stephane Ayache, Yağmur Güçlütürk, and Umut Güçlü

1 Introduction

Dealing with missing and incomplete information is a very relevant problem common for several tasks and scenarios in computer vision and pattern recognition. Together with related tasks like denoising, deblurring, super-resolution, enhancement, etc. inpainting aims at *generating* visual information with models that usually exploit the context of a corrupted visual input. This is a very complex task, because the goal is to produce visual content that is satisfactory and attractive to the human visual system. With the rapid progress of deep learning, impressive solutions for all of these tasks have been developed recently (see e.g., [1]). In order to keep track of all of this progress we edited this special issue focusing on Image and Video Inpainting and Denoising and related tasks.

The scope of the issue comprised all aspects of computer vision and pattern recognition devoted to image and video inpainting, including related tasks like denoising, deblurring, sampling, super-resolutkon enhancement, restoration, hallucination, etc. The special issue was associated to the 2018 Chalearn Looking at People Satellite ECCV Workshop[1] and the 2018 ChaLearn Challenges on Image and Video Inpainting.[2] However, the call for papers was open to the public. A dozen of submissions were received, and every paper was subject to the standard TPAMI review process.

1. http://chalearnlap.cvc.uab.es/workshop/29/description/
2. http://chalearnlap.cvc.uab.es/challenge/26/description/

- *S. Escalera is with the University of Barcelona and Computer Vision Center, 08007 Barcelona, Spain. E-mail: sergio@maia.ub.es.*
- *H.J. Escalante is with Instituto Nacional de Astrofísica, Óptica y Eectrónica, Puebla, Mexico, and also with Computer Science Department, CINVESTAV Zacatenco, Mexico City 07360, Mexico. E-mail: hugojair@inaoep.mx.*
- *X. Baró is with Universitat Oberta de Catalunya and Computer Vision Center, 08018 Barcelona, Spain. E-mail: xbaro@uoc.edu.*
- *I. Guyon is with UPSud/INRIA Université Paris-Saclay, 91190 Essonne, France, and also with Président of ChaLearn, Berkeley, CA. E-mail: guyon@chalearn.org.*
- *M. Madadi is with Universitat de Barcelona and Computer Vision Center, 08007 Barcelona, Spain. E-mail: meysam.madadi@gmail.com.*
- *J. Wan is with National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, and the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China. E-mail: jun.wan@ia.ac.cn.*
- *S. Ayache is with Laboratoire d'Informatique et Systèmes, Aix-Marseille Université, 13007 Marseille, France. E-mail: stephane.ayache@lis-lab.fr.*
- *Y. Güçlütürk and U. Güçlü are with Donders Institute for Brain, Cognition and Behaviour, Radboud University, 6525 XZ Nijmegen, the Netherlands. E-mail: {ygucluturk, umuguc}@gmail.com.*

Digital Object Identifier no. 10.1109/TPAMI.2020.2971291

In the remainder of this note we briefly summarize the contributions of the articles included in the issue and identify trends and research opportunities on the topic.

2 The Image and Video Inpainting and Denoising Special Issue

This special issue is formed by 3 articles of outstanding quality that together comprise a snapshot of cutting edge techniques on inpainting and denoising of visual information. Table 1 summarizes the main characteristics of the accepted papers.

The contributions focused on different but very related tasks, and the three papers proposed end to end deep learning models with outstanding methodological contributions. On the one hand, He *et al.* dealt with the multimodal face completion problem: synthesizing an RGB image depicting a face from an NIR input [2]. They introduced a cross-spectral model formed by a NIR-RGB GAN synthesizer followed by a pose correction module. Throughout an extensive experimental evaluation they showed their model is robust against pose variations and achieved state of the art performance on face recognition.

The remaining papers that form the special issue dealt with video-related tasks. Kim *et al.* describe two models for addressing the video de-captioning and object removal from videos tasks [6]. An encoder-decoder model taking as input consecutive frames that are aggregated and decoded is proposed. The model is enhanced with recurrent feedback to enforce temporal consistency. With a variant of this model, the authors obtained the best results in the track 2 of the 2018 ChaLearn Challenge on Image and Video Inpainting [7]. Additionally, they show the generality of the proposed model by reporting experiments on generic object removal from videos, where another variant of the model showed outstanding performance.

Szeto *et al.* also worked with video, however, they approached a very novel inpainting task: video frame inpainting, that is, the problem of inferring missing frames in a video [10]. A two stage deep learning model is proposed. First, a bidirectional convolutional LSTM model conditioned on the preceding and following frames makes two intermediate frame predictions. Then an interpolation model aims at blending both predictions by taking into account temporal and hidden activations information to generate a single predicted

TABLE 1
Overview of Articles in the Special Issue on Imagen and Video Inpainting and Denosing

Ref.	Task / Model	Model	Dataset
[2]	NIR-Visible Face Completion, Face Recognition	End-to-end deep neural network combining a texture inpainting model (NIR → Visible) and a pose correction model	CASIA NIR-VIS 2.0 [3], BUAA-VisNir face database [4], Oulu-CASIA NIR-VIS database [5]
[6]	Video inpainting, de-captioning, object removal from video	Recurrent temporal model for frame aggregation based on a encoder-decoder architecture	ChaLearn video De-Captioning [7]. Youtube-VOS [8], DAVIS [9]
[10]	Video frame inpainting	Bidirectional (convolutional LSTM-based encoder-decoder) prediction model and a temporally aware frame interpolation model	KTH Actions [11], UCF-101 [12], HMDB-51 [13]

frame. Impressive results are obtained in human action data sets commonly used for frame prediction and related tasks.

3 Discussion

The special issue is a compilation of cutting-edge research on inpainting and denoising of visual information. Although the topic is very popular among the computer vision and pattern recognition communities, we received a dozen of high quality submissions and after a compelling reviewing procedure three papers were accepted. The three manuscripts deal with very novel tasks, from cross-spectral image synthesis, to video de-captioning and to frame inpainting in videos. Novel deep learning architectures were proposed and every component of the architecture was evaluated, state of the art performance is reported in each of these publications.

Based on the accepted papers and on the received submissions we can outline the following conclusions:

- Visual inpainting and denosing are two very fast moving research fields, with new models and impressive results being reported regularly. This fact makes researchers to focus on fast publication forums.
- Inpainting is a broad reaching methodology having impact into a number of tasks and applications. Three novel tasks were approached in the accepted papers, they surely will motivate further research in the forthcoming years.
- Deep learning with GAN-based learning is a consolidated methodology within inpaiting of visual information. The deep learning architectures that were proposed were sound and designated to solve very specific problems found in the approached tasks. It is difficult to think that automatic solutions (e.g., AutoML [14]) may work reasonably well for this domain.
- As deep learning solutions dominate the field, the creation of large scale and well curated resources will be critical for the next few years in this line of research.

Acknowledgments

This work was supported in part by the Spanish projects TIN2015-66951-C2-2-R and TIN2016-74946-P (MINECO/FEDER, UE) and CERCA Programme/Generalitat de Catalunya and by INAOE, CONACyT-Mexico under grant CB-A1-S-26314. This work was also supported in part by ICREA under the ICREA Academia programme. The authors would like to thank ChaLearn Looking at People sponsors for their support, including Microsoft Research, Google, NVIDIA Coorporation, Amazon, Facebook, and Disney Research.

References

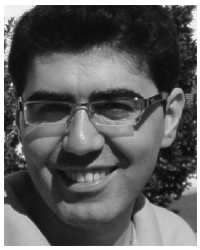
[1] S. Escalera, S. Ayache, J. Wan, M. Madadi, U. Güçlü, and X. Baró, *Inpainting and Denoising Challenges*. Berlin, Germany: Springer, 2019.
[2] R. He, J. Cao, L. Song, Z. Sun, and T. Tan, "Adversarial cross-spectral face completion for NIR-VIS face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2019.2961900.
[3] S. Z. Li, D. Yi, Z. Lei, and S. Liao, "The CASIA NIR-VIS 2.0 face database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 348–353. [Online]. Available: https://doi.org/10.1109/CVPRW.2013.59
[4] D. Huang, J. Sun, and Y. Wang, "The BUAA-VisNir face database instructions," Beihang Univ., Beijing, China, Tech. Rep. IRIP- TR-12-FR-001, 2012.
[5] J. Chen, D. Yi, J. Yang, G. Zhao, S. Z. Li, and M. Pietikäinen, "Learning mappings for face synthesis from near infrared to visual light images," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 156–163.
[6] D. Kim, S. Woo, J.-Y. Lee, and I. S. Kweon, "Recurrent temporal aggregation framework for deep video inpainting," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2019.2958083.
[7] S. Escalera *et al.*, "ChaLearn looking at people: Inpainting and denoising challenges," in *Proc. Inpainting Denoising Challenges*, 2019, pp. 23–44.
[8] N. Xu *et al.*, "YouTube-VOS: Sequence-to-sequence video object segmentation," in *Proc. 15th Eur. Conf. Comput. Vis.*, 2018, pp. 603–619. [Online]. Available: https://doi.org/10.1007/978-3-030-01228-1_36
[9] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. V. Gool, M. H. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 724–732.
[10] R. Szeto, X. Sun, K. Lu, and J. Corso, "A temporally-aware interpolation network for video frame inpainting," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2019.2951667.
[11] C. Schüldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. 17th Int. Conf. Pattern Recognit.*, 2004, pp. 32–36. [Online]. Available: https://doi.org/10.1109/ICPR.2004.1334462
[12] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *CoRR*, 2012. [Online]. Available: http://arxiv.org/abs/1212.0402
[13] H. Kuehne, H. Jhuang, E. Garrote, T. A. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 2556–2563.
[14] Z. Liu *et al.*, "Autocv challenge design and baseline results," in *Proc. La Conf. sur l'Apprentissage Automatique*, HAL archive, 2019.

Sergio Escalera received the PhD degree on multiclass visual categorization systems from Computer Vision Center, UAB. He received the 2008 Best Thesis Award. He leads the Human Pose Recovery and Behavior Analysis Group at Universitat de Barcelona and Computer Vision Center. He is currently an associate professor with the Department of Mathematics and Informatics, Universitat de Barcelona, Barcelona, Spain. He is also a member of the Computer Vision Center at Campus UAB. He is vice-president of ChaLearn Challenges in Machine Learning and chair of IAPR TC-12: Multimedia and visual information systems. He is ICREA Academia and member of the European Laboratory for Learning and Intelligent Systems ELLIS. His research interests include automatic analysis of humans from visual and multi-modal data, with special interest in inclusive, transparent, and fair affective computing and people characterization: personality and psychological profile computing.

Hugo Jair Escalante is researcher scientist at Instituto Nacional de Astrofisica, Optica y Electronica, INAOE, Mexico and a director of ChaLearn, a nonprofit dedicated to organizing challenges, since 2011. He has been involved in the organization of several challenges in computer vision and automatic machine learning. He is a reviewer at the *Journal of Machine Learning Research*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, and has served as co-editor of special issues in the *International Journal of Computer Vision*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, and the *Theoretical and Applied Climatology*. He has served as competition chair and area chair of venues like NeurIPS, PAKDD, IJCNN, among others. He serving as area chair for NIPS 2016, and has been member of the program committee of venues like CVPR, ICCV, ECCV, ICML, NIPS, IJCNN.

Xavier Baró received the BS degree in computer science from the UAB, in 2003, the MS degree in computer science from UAB, in 2005, and the PhD degree in computer engineering, in 2009. He is currently an associate professor and researcher with the Computer Science, Multimedia and Telecommunications Department, Universitat Oberta de Catalunya (UOC). From 2015, he is the director of the Computer Vision Master degree at UOC. He is co-founder of the Scene Understanding and Artificial Intelligence (SUNAI) group, and collaborates with the Computer Vision Center of the UAB, as member of the Human Pose Recovery and Behavior Analysis (HUPBA) group. His research interests are related to machine learning, evolutionary computation, and statistical pattern recognition, specially their applications to the field of Looking at People.

Isabelle Guyon is currently a chair professor in Big Data with the Université Paris-Saclay (UPSud/INRIA), specialized in statistical data analysis, pattern recognition and machine learning. Her areas of expertise include computer vision and bioinformatics. Her recent interest is in applications of machine learning to the discovery of causal relationships. Prior to joining Paris-Saclay she worked as an independent consultant and was a researcher at AT&T Bell Laboratories, where she pioneered applications of neural networks to pen computer interfaces (with collaborators including Yann LeCun and Yoshua Bengio) and co-invented with Bernhard Boser and Vladimir Vapnik Support Vector Machines (SVM), which became a textbook machine learning method. She organizes challenges in Machine Learning since 2003 supported by the EU network Pascal2, NSF, and DARPA, with prizes sponsored by Microsoft, Google, Facebook, Amazon, Disney Research, and Texas Instrument. She is action editor of the *Journal of Machine Learning Research*, program chair of the NIPS 2016 conference, and general chair of the NIPS 2017 conference.



Meysam Madadi received the MS and PhD degrees in computer vision from the Universitat Autònoma de Barcelona (UAB), in 2013 and 2017, respectively. He is currently a postdoc researcher at Computer Vision Center (CVC), UAB. He has been a member of Human Pose Recovery and Behavior Analysis (HUPBA) group since 2012. He collaborated in ChalearnLAP ECCV2014, NIPS2016, CVPR2017, ICCV17, ECCV18, and CVPR19.

Jun Wan (Senior Member, IEEE) received the BS degree from the China University of Geosciences, Beijing, China, in 2008, and the PhD degree from the Institute of Information Science, Beijing Jiaotong University, Beijing, China, in 2015. Since January 2015, he has been a faculty member with the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Science (CASIA), China, where he currently serves as an associate professor. He is a director of Chalearn Challenges. He has published more than 45 research papers and has been guest editor at the *IEEE Transactions on Pattern Analysis and Machine Intelligence*, the *Journal of Machine Vision and Applications* and the *Entropy*. His main research interests include computer vision, machine learning, especially for face and pedestrian analysis (such as attribute analysis, face anti-spoofing detection), gesture and sign language recognition. He has published papers in top journals and conferences, such as the *Journal of Machine Learning Research*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Transactions on Image Processing*, *IEEE Transactions on Multimedia*, *IEEE Transactions on Cybernetics*, *ACM Transactions on Multimedia Computing, Communications, and Applications*, the *PR Journal*, the *Computer Vision and Image Understanding*, CVPR, AAAI, IJCAI. He has served as the reviewer on several journals and conferences, such as the *Journal of Machine Learning Research*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Transactions on Image Processing*, *IEEE Transactions on Multimedia*, *IEEE Transactions on Systems, Man and Cybernetics*, *PR Journal*, CVPR, ICCV, ECCV, AAAI, and ICRA.

Stephane Ayache received the PhD degree in computer science from Institut National Polytechnic de Grenoble (INPG), France, in 2007, he is currently associate professor at AMU/LIS since 2008. His research interests include computer vision, image/video content understanding and indexing, machine learning with a focus on multimodal analysis and multiview learning, transfer learning, and deep learning.

Yağmur Güçlütürk received the BIT degree in artificial intelligence from the Multimedia University, Cyberjaya, Malaysia, and the MSc and PhD degrees in cognitive neuroscience from the Radboud University, Nijmegen, Netherlands. She is an assistant professor with the Radboud University, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, Netherlands, where she is currently working on neuroprosthetics and AI as a member of the Artificial Cognitive Systems Lab and the Neuronal Stimulation for Recovery of Function Consortium. Her graduate studies and research assistantship were funded by the prestigious fellowships from the Netherlands Universities Foundation for International Cooperation and the Royal Netherlands Academy of Arts and Sciences, respectively.

Umut Güçlü received the bachelor's degree in artificial intelligence, as well as additional undergraduate studies in natural sciences and mathematics and the master's, postdoctoral, and doctoral degrees in cognitive neuroscience. He is currently an assistant professor at Radboud University, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, Netherlands, and a member of the Artificial Cognitive Systems Lab, where he works on combining deep learning and neural coding to systematically investigate the cognitive algorithms in vivo with neuroimaging and implement them in silico with artificial neural networks. He also works on decoding the mental lexicon in the Language in Interaction consortium and restoring the visual function in the Neuronal Stimulation for Recovery of Function consortium.